(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
3. Mai 2001 (03.05.2001)

PCT

(10) Internationale Veröffentlichungsnummer
WO 01/31434 A2

(51) Internationale Patentklassifikation[7]: G06F 3/16

(21) Internationales Aktenzeichen: PCT/DE00/03753

(22) Internationales Anmeldedatum:
24. Oktober 2000 (24.10.2000)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
199 52 051.8 28. Oktober 1999 (28.10.1999) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: SIEMENS AKTIENGESELLSCHAFT [DE/DE]; Wittelsbacher Platz 2, 80333 München (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder *(nur für US)*: HOLZAPFEL, Martin [DE/DE]; Stabelerstrasse 13, 80933 München (DE). ERDEM, Caglayan [DE/DE]; Barlachstrasse 6, 80804 München (DE).

(74) Gemeinsamer Vertreter: SIEMENS AKTIENGESELLSCHAFT; Postfach 22 16 34, 80506 München (DE).

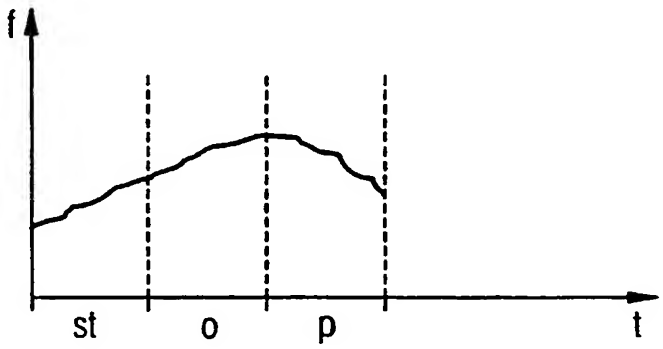
(81) Bestimmungsstaaten *(national)*: JP, US.

(84) Bestimmungsstaaten *(regional)*: europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

*[Fortsetzung auf der nächsten Seite]*

(54) Title: METHOD FOR DETECTING THE TIME SEQUENCES OF A FUNDAMENTAL FREQUENCY OF AN AUDIO-RESPONSE UNIT TO BE SYNTHESISED

(54) Bezeichnung: VERFAHREN ZUM BESTIMMEN DES ZEITLICHEN VERLAUFS EINER GRUNDFREQUENZ EINER ZU SYNTHETISIERENDEN SPRACHAUSGABE




(57) Abstract: The invention relates to a method for detecting the time sequences of a fundamental frequency of an audio-response unit to be synthesised. The invention is characterised in that input macro segments of the fundamental frequency are detected by means of a neuronal network and are reproduced by means of fundamental frequency sequences that are stored in a data base. According to the inventive method, the fundamental frequency is produced based on a greater text section which is analysed by means of the neuronal network. Microstructures are transferred from the data base to the fundamental frequency. The thus produced fundamental frequency is optimised in the macro and microstructure thereof. An extremely natural sound is thus obtained.

(57) Zusammenfassung: Die Erfindung betrifft ein Verfahren zum Bestimmen des zeitlichen Verlaufs einer Grundfrequenz einer zu synthetisierenden Sprachausgabe. Die Erfindung zeichnet sich dadurch aus, daß Vorgabemakrosegmente der Grundfrequenz mittels eines neuronalen Netzwerkes bestimmt werden, und diese Vorgabemakrosegmente mittels in einer Datenbasis gespeicherten Grundfrequenzsequenzen nachgebildet werden. Durch das erfindungsgemäße Verfahren wird die Grundfrequenz auf Grundlage eines größeren Textabschnittes, der mittels des neuronalen Netzwerkes analysiert wird, erzeugt, wobei aus der Datenbasis Mikrostrukturen in der Grundfrequenz aufgenommen werden. Die derart gebildete Grundfrequenz ist somit bezüglich ihrer Makro- als auch ihrer Mikrostruktur optimiert. Hierdurch wird ein äußerst natürlicher Klang erzielt.

WO 01/31434 A2

**Veröffentlicht:**

— *Ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts.*

*Zur Erklärung der Zweibuchstaben-Codes, und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.*

Beschreibung

Verfahren zum Bestimmen des zeitlichen Verlaufs einer Grundfrequenz einer zu synthetisierenden Sprachausgabe

Die Erfindung betrifft ein Verfahren zum Bestimmen des zeitlichen Verlaufs einer Grundfrequenz einer zu synthetisierenden Sprachausgabe.

Auf der Konferenz ICASSP 97, in München, ist unter dem Titel „Recent Improvements on Microsoft's Trainable Text-to-Speech System-Whistler", X. Huang et al, ein Verfahren zum Synthetisieren von Sprache aus einem Text vorgestellt worden, das vollständig trainierbar ist und die Prosodie eines Textes anhand von in einer Datenbank gespeicherten Prosodiemustern zusammenstellt und erzeugt. Die Prosodie eines Textes wird im wesentlichen durch die Grundfrequenz festgelegt, weshalb dieses bekannte Verfahren auch als Verfahren zur Erzeugung einer Grundfrequenz auf Grundlage entsprechender in einer Datenbank gespeicherter Muster betrachtet werden kann. Zur Erzielung einer möglichst natürlichen Sprachweise sind aufwendige Korrekturverfahren vorgesehen, die die Kontur der Grundfrequenz interpolieren, glätten und korrigieren.

Auf der ICASSP 98, in Seattle, ist unter dem Titel „Optimization of a Neural Network for Speaker and Task Dependent $F_0$-Generation", Ralf Haury et al. ein weiteres Verfahren zum Erzeugen einer synthetischen Sprachausgabe aus einem Text vorgestellt worden. Dieses bekannte Verfahren verwendet zur Erzeugung der Grundfrequenz anstelle einer Datenbank mit Mustern ein neuronales Netzwerk, mit dem der zeitliche Verlauf der Grundfrequenz für die Sprachausgabe festgelegt wird.

Mit den oben beschriebenen Verfahren soll eine Sprachausgabe geschaffen werden, die keinen metallischen, mechanischen und unnatürlichen Klang besitzt, wie es von herkömmlichen Sprachsynthesesystemen bekannt ist. Diese Verfahren stellen eine

deutliche Verbesserung gegenüber den herkömmlichen Sprachsynthesesystemen dar. Es bestehen dennoch erhebliche klangliche Unterschiede zwischen der auf diesen Verfahren beruhenden Sprachausgabe und einer menschlichen Stimme.

Insbesondere wird bei einer Sprachsynthese, bei der die Grundfrequenz aus einzelnen Grundfrequenzmustern zusammengesetzt wird, nach wie vor ein metallischer, mechanischer Klang erzeugt, der deutlich von einer natürlichen Stimme unterschieden werden kann. Wird die Grundfrequenz hingegen mit einem neuronalen Netzwerk festgelegt, klingt die Stimme zwar natürlicher, aber ist etwas dumpf.

Der Erfindung liegt deshalb die Aufgabe zugrunde, ein Verfahren zum Bestimmen des zeitlichen Verlaufs einer Grundfrequenz einer zu synthetisierenden Sprachausgabe zu schaffen, die der Sprachausgabe einen natürlichen, einer menschlichen Stimme sehr ähnlichen Klang verleiht.

Die Aufgabe wird durch ein Verfahren mit den Merkmalen des Anspruchs 1 gelöst. Vorteilhafte Ausgestaltungen sind in den Unteransprüchen angegeben.

Das erfindungsgemäße Verfahren zum Bestimmen des zeitlichen Verlaufs einer Grundfrequenz einer zu synthetisierenden Sprachausgabe umfaßt folgende Schritte:

Bestimmen von Vorgabemakrosegmenten der Grundfrequenz mittels eines neuronalen Netzwerkes, und

Bestimmen von Mikrosegmenten mittels in einer Datenbasis gespeicherten Grundfrequenzsequenzen, wobei die Grundfrequenzsequenzen derart aus der Datenbasis ausgewählt werden, daß durch die aufeinanderfolgenden Grundfrequenzsequenzen das jeweilige Vorgabemakrosegment mit möglichst geringer Abweichung nachgebildet wird.

Der vorliegenden Erfindung liegt die Erkenntnis zugrunde, daß die Bestimmung des Verlaufs einer Grundfrequenz mittels eines neuronalen Netzwerkes die Makrostruktur des zeitlichen Verlaufs einer Grundfrequenz sehr ähnlich zu dem Verlauf der Grundfrequenz einer natürlichen Sprache erzeugt, und die in einer Datenbasis gespeicherten Grundfrequenzsequenzen sehr ähnlich die Mikrostruktur der Grundfrequenz einer natürlichen Sprache wiedergeben. Durch die erfindungsgemäße Kombination wird somit eine optimale Bestimmung des Verlaufs der Grundfrequenz erzielt, die sowohl in der Makrostruktur als auch in der Mikrostruktur der der natürlichen Sprache wesentlich ähnlicher ist, als bei einer mit den bisher bekannten Verfahren erzeugten Grundfrequenz. Hierdurch wird eine beträchtliche Annäherung der synthetischen Sprachausgabe an eine natürliche Sprache erzielt. Die hierdurch erzeugte synthetische Sprache ist der natürlichen Sprache sehr ähnlich und kann kaum von dieser unterschieden werden.

Vorzugsweise wird die Abweichung zwischen dem Nachbildungsmakrosegment und dem Vorgabemakrosegment mittels einer Kostenfunktion ermittelt, die derart gewichtet ist, daß bei geringen Abweichungen von der Grundfrequenz des Vorgabemakrosegments lediglich eine kleine Abweichung ermittelt wird, wobei bei Überschreitung vorbestimmter Grenzfrequenzdifferenzen die ermittelten Abweichungen stark bis zum Erreichen eines Sättigungswertes ansteigen. Dies bedeutet, daß alle Grundfrequenzsequenzen, die innerhalb des Bereiches der Grenzfrequenzen liegen, eine sinnvolle Auswahl zur Nachbildung des Vorgabemakrosegments darstellen und die Grundfrequenzsequenzen, die außerhalb des Bereiches der Grenzfrequenzdifferenzen liegen, als wesentlich ungeeigneter zur Nachbildung des Vorgabemakrosegments bewertet werden. Diese Nichtlinearität bildet das nichtlineare Verhalten des menschlichen Gehörs ab.

Nach einer weiteren bevorzugten Ausführungsform der Erfindung werden Abweichungen desto schwächer gewichtet, je näher sie am Rand einer Silbe angeordnet sind.

Die Nachbildung des Vorgabemakrosegments erfolgt vorzugsweise durch Erzeugung mehrerer Grundfrequenzsequenzen für jeweils eine mikroprosodische Einheit, wobei Kombinationen von Grundfrequenzsequenzen sowohl bezüglich der Abweichung vom Vorgabemakrosegment als auch bezüglich einer paarweisen Abstimmung bewertet werden. In Abhängigkeit des Ergebnisses dieser beiden Bewertungen (Abweichung vom Vorgabemakrosegment, Abstimmung zwischen benachbarten Grundfrequenzsequenzen) wird dann eine entsprechende Auswahl einer Kombination von Grundfrequenzsequenzen getroffen.

Mit dieser paarweisen Abstimmung werden insbesondere die Übergänge zwischen benachbarten Grundfrequenzsequenzen bewertet, wobei hier größere Sprünge vermieden werden sollen. Nach einer bevorzugten Ausführungsform der Erfindung werden diese paarweisen Abstimmungen der Grundfrequenzsequenzen innerhalb einer Silbe stärker gewichtet als am Randbereich der Silbe. Der Silbenkern ist im Deutschen maßgeblich für den Höreindruck.

Das erfindungsgemäße Verfahren wird nachfolgend anhand eines in der Zeichnung dargestellten Ausführungsbeispieles näher erläutert. In den Zeichnungen zeigen schematisch:

Fig. 1a bis 1d den Aufbau und das Zusammensetzen des zeitlichen Verlaufes einer Grundfrequenz in vier Schritten,

Fig. 2 eine Funktion zur Gewichtung einer Kostenfunktion zur Bestimmung der Abweichung zwischen einem Nachbildungsmakrosegment und einem Vorgabemakrosegment,

Fig. 3 den Verlauf einer aus mehreren Makrosegmenten bestehenden Grundfrequenz,

Fig. 4 schematisch vereinfacht den Aufbau eines neuronalen Netzwerkes,

Fig. 5 das erfindungsgemäße Verfahren in einem Flußdiagramm, und

Fig. 6 ein Verfahren zum Synthetisieren von Sprache, daß auf dem erfindungsgemäßen Verfahren beruht.

In Fig. 6 ist ein Verfahren zum Synthetisieren von Sprache, bei dem ein Text in eine Folge akustischer Signale gewandelt wird, in einem Flußdiagramm dargestellt.

Dieses Verfahren ist in Form eines Computerprogrammes realisiert, das mit einem Schritt S1 gestartet wird.

Im Schritt S2 wird ein Text eingegeben, der in Form einer elektronisch lesbaren Textdatei vorliegt.

Im folgenden Schritt S3 wird eine Folge von Phonemen, das heißt eine Lautfolge, erstellt, wobei den einzelnen Graphemen des Textes, das sind jeweils einzelne oder mehrere Buchstaben, denen jeweils ein Phonem zugeordnet ist, ermittelt werden. Es werden dann die den einzelnen Graphemen zugeordneten Phoneme bestimmt, wodurch die Phonemfolge festgelegt ist.

Im Schritt S4 wird eine Betonungsstruktur bestimmt, das heißt es wird bestimmt, wie stark die einzelnen Phoneme betont werden sollen.

Die Betonungsstruktur ist in Fig. 1a mittels eines Zeitstrahles anhand des Wortes „stop" dargestellt. Demgemäß sind dem Graphem „st" die Betonungsstufe 1, dem Graphem „o" die Betonungsstufe 0,3 und dem Graphem „p" die Betonungsstufe 0,5 zugeordnet worden.

Nachfolgend wird die Dauer der einzelnen Phoneme bestimmt (S5).

Im Schritt S6 wird der zeitliche Verlauf der Grundfrequenz bestimmt, was unten näher ausgeführt ist.

Nachdem die Phonemfolge und die Grundfrequenz festgelegt sind, kann eine Wave-Datei auf Grundlage der Phoneme und der Grundfrequenz erzeugt werden (S7).

Die Wave-Datei wird mittels einer akustischen Ausgabeeinheit und einem Lautsprecher in akustische Signale umgesetzt (S8), womit die Sprachausgabe beendet ist (S9).

Erfindungsgemäß wird der zeitliche Verlauf der Grundfrequenz der zu synthetisierenden Sprachausgabe mittels eines neuronalen Netzwerkes in Kombination mit in einer Datenbasis gespeicherten Grundfrequenzsequenzen erzeugt.

Das Verfahren, das dem Schritt S6 aus Fig. 6 entspricht, ist ausführlicher in Fig. 5 in einem Flußdiagramm dargestellt.

Dieses Verfahren zum Bestimmen des zeitlichen Verlaufs der Grundfrequenz ist ein Unterprogramm zu dem in Fig. 6 gezeigtem Programm. Das Unterprogramm wird mit dem Schritt S10 gestartet.

Mit dem Schritt S11 wird ein Vorgabemakrosegment der Grundfrequenz mittels eines neuronalen Netzwerkes bestimmt. Ein derartiges neuronales Netzwerk ist schematisch vereinfacht in Fig. 4 gezeigt. Das neuronale Netzwerk weist an einer Eingabeschicht I Knoten zur Eingabe einer phonetisch linguistischen Einheit PE des zu synthetisierenden Textes und eines Kontextes Kl, Kr links und rechts von der phonetisch linguistischen Einheit auf. Die phonetisch linguistische Einheit besteht z.B. aus einer Phrase, einem Wort oder einer Silbe des zu synthetisierenden Textes, zu der das Vorgabemakroseg-

ment der Grundfrequenz bestimmt werden soll. Der linke Kontext Kl und der rechte Kontext Kr stellen jeweils einen Textabschnitt links und rechts der phonetischen linguistischen Einheit PE dar. Die mit der phonetischen Einheit eingegebenen Daten umfassen die entsprechende Phonemfolge, Betonungsstruktur und die Lautdauer der einzelnen Phoneme. Die mit dem linken bzw. rechten Kontext eingegebenen Informationen umfassen zumindest die Phonemfolge, wobei es zweckmäßig sein kann, auch die Betonungsstruktur und/oder die Lautdauer mit einzugeben. Die Länge des linken und rechten Kontextes kann der Länge der phonetisch linguistischen Einheit PE entsprechen, also wiederum eine Phrase, ein Wort oder eine Silbe sein. Es kann jedoch auch zweckmäßig sein, einen längeren Kontext von z.B. zwei oder drei Wörtern als linken oder rechten Kontext vorzusehen. Diese Eingaben Kl, PE und Kr werden in einer versteckten Schicht VS verarbeitet und an einer Ausgabeschicht O als Vorgabemakrosegment VG der Grundfrequenz ausgegeben.

In Fig. 1b ist eine solche Vorgabemakrosegment für das Wort „stop" dargestellt. Dieses Vorgabemakrosegment besitzt einen typischen dreiecksförmigen Verlauf, der zunächst mit einem Anstieg beginnt und mit einem etwas kürzeren Abfall endet.

Nach der Bestimmung eines Vorgabemakrosegmentes der Grundfrequenz werden in den Schritten S12 und S13 die dem Vorgabemakrosegment entsprechenden Mikrosegmente bestimmt.

Im Schritt S12 werden aus einer Datenbasis, in der Graphemen zugeordnete Grundfrequenzsequenzen gespeichert sind, ausgelesen, wobei in der Regel für jedes Graphem eine Vielzahl von Grundfrequenzsequenzen vorliegen. In Fig. 1c sind derartige Grundfrequenzsequenzen für die Grapheme „st", „o" und „p" schematisch dargestellt, wobei zur zeichnerischen Vereinfachung lediglich eine geringe Anzahl von Grundfrequenzsequenzen gezeigt sind.

Diese Grundfrequenzsequenzen können grundsätzlich beliebig miteinander kombiniert werden. Die möglichen Kombinationen dieser Grundfrequenzsequenzen werden mittels einer Kostenfunktion bewertet. Dieser Verfahrensschritt wird mittels des Viterbi-Algorhithmus ausgeführt.

Für jede Kombination von Grundfrequenzsequenzen, die für jedes Phonem eine Grundfrequenzsequenz aufweist, wird ein Kostenfaktor Kf mittels folgender Kostenfunktion berechnet:

$$Kf = \sum_{j=1}^{j=l} lok(f_{ij}) + Verk(f_{ij}, f_{n,j+1})$$

Die Kostenfunktion ist eine Summe von j=1 bis l, wobei j der Zähler der Phoneme ist und l die Gesamtzahl aller Phoneme darstellt. Die Kostenfunktion weist zwei Terme auf, eine lokale Kostenfunktion lok ($k_{ij}$)und eine Verknüpfungskostenfunktion Ver($k_{ij}$, $k_n$, j+1). Mit der lokalen Kostenfunktion wird die Abweichung der i-ten Grundfrequenzsequenz des j-ten Phonems vom Vorgabemakrosegment bewertet. Mit der Verknüpfungskostenfunktion wird die Abstimmung zwischen der i-ten Grundfrequenz des j-ten Phonems mit der n-ten Grundfrequenzsequenz des j+1-ten Phonems bewertet.

Die lokale Kostenfunktion weist beispielsweise folgende Form auf:

$$lok(f_{ij}) = \int_{ta}^{te} (f_V(t) - f_{ij}(t))^2 dt$$

Die lokale Kostenfunktion ist somit ein Integral über den Zeitbereich des Beginns ta eines Phonems bis zum Ende te des Phonems über das Quadrat der Differenz der durch das Vorgabemakrosegment vorgegebenen Grundfrequenz $f_V$ und der i-ten Grundfrequenzsequenz des j-ten Phonems.

Diese lokale Kostenfunktion ermittelt somit einen positiven Wert der Abweichung zwischen der jeweiligen Grundfrequenzsequenz und der Grundfrequenz des Vorgabemakrosegments. Zudem ist diese Kostenfunktion sehr einfach realisierbar und erzeugt durch die parabolische Eigenschaft eine Bewertung, die der des menschlichen Gehörs ähnelt, da kleinere Abweichungen um die Vorgabeseqeunz $f_V$ gering bewertet werden, wohingegen größere Abweichungen progressiv bewertet werden.

Nach einer bevorzugten Ausführungsform wird die lokale Kostenfunktion mit einem Gewichtungsterm versehen, der zu dem in Fig. 2 dargestellten Funktionsverlauf führt. Das Diagramm aus Fig. 2 zeigt den Wert der lokalen Kostenfunktion lok ($f_{ij}$) in Abhängigkeit vom Logorhitmus der Frequenz $f_{ij}$ der i-ten Grundfrequenzsequenz des j-ten Phonems. Dem Diagramm kann man entnehmen, daß Abweichungen von der Vorgabefrequenz $f_V$ innerhalb bestimmter Grenzfrequenzen GF1, GF2 nur gering bewertet werden, wobei eine weitere Abweichung einen stark zunehmenden Anstieg bis zu einem Schwellwert SW bewirkt. Eine derartige Gewichtung entspricht dem menschlichen Gehör, das geringe Frequenzabweichungen kaum wahrnimmt aber ab gewissen Frequenzdifferenzen dies als deutlichen Unterschied registriert.

Mit der Verknüpfungskostenfunktion wird bewertet, wie gut zwei aufeinanderfolgende Grundfrequenzsequenzen aufeinander abgestimmt sind. Insbesondere wird hierbei die Frequenzdifferenz an der Verbindungsstelle der beiden Grundfrequenzsequenzen bewertet, wobei je größer die Differenz am Ende der vorhergehenden Grundfrequenzsequenz zur Frequenz am Anfang der nachfolgenden Grundfrequenzsequenzen ist, desto größer ist der Ausgabewert der Verknüpfungskostenfunktion. Hierbei können jedoch noch weitere Parameter berücksichtigt werden, die z.B. die Stetigkeit des Überganges oder dergleichen, wiedergeben.

Bei einer bevorzugten Ausführungsform der Erfindung wird der Ausgabewert der Verknüpfungskostenfunktion umso schwächer gewichtet, je näher die jeweilige Verbindungsstelle zweier benachbarter Grundfrequenzsequenzen am Rand einer Silbe angeordnet ist. Dies entspricht dem menschlichen Gehör, das akustische Signale am Rande einer Silbe weniger intensiv analysiert als im mittleren Bereich der Silbe. Eine derartige Gewichtung wird auch als perzeptiv dominant bezeichnet.

Gemäß obiger Kostenfunktion Kf werden für jede Kombination von Grundfrequenzsequenzen der Phoneme einer linguistischen Einheit, für die ein Vorgabemakrosegment bestimmt worden ist, die Werte der lokalen Kostenfunktion und der Verknüpfungskostenfunktion aller Grundfrequenzsequenzen ermittelt und summiert. Aus der Menge der Kombinationen der Grundfrequenzsequenzen wird diejenige Kombination ausgewählt, für die die Kostenfunktion Kf den kleinsten Wert ergeben hat, da diese Kombination von Grundfrequenzsequenzen einen Grundfrequenzverlauf für die entsprechende linguistische Einheit bildet, der als Nachbildungsmakrosegment bezeichnet wird und dem Vorgabemakrosegment sehr ähnlich ist.

Mit dem erfindungsgemäßen Verfahren werden somit an die mittels des neuronalen Netzwerkes erzeugten Vorgabemakrosegemente der Grundfrequenz angepaßte Grundfrequenzverläufe mittels einzelner in einer Datenbasis gespeicherten Grundfrequenzsequenzen erzeugt. Hierdurch wird eine sehr natürliche Makrostruktur sichergestellt, die zudem auch die detailgenaue Mikrostruktur der Grundfrequenzsequenzen besitzt.

Ein derartiges Nachbildungsmakrosegment für das Wort „stop" ist in Fig. 1d gezeigt.

Nachdem im Schritt S13 die Auswahl der Kombinationen von Grundfrequenzsequenzen zur Nachbildung des Vorgabemakrosegments abgeschlossen ist, wird im Schritt S14 geprüft, ob für eine weitere phonetische linguistische Einheit ein weiterer

zeitlicher Verlauf der Grundfrequenz erzeugt werden muß. Ergibt diese Abfrage im Schritt S14 ein „ja", springt der Programmablauf auf den Schritt S11 zurück, andernfalls verzweigt der Programmablauf auf den Schritt S15, mit dem die einzelnen Nachbildungsmakrosegmente der Grundfrequenz zusammengesetzt werden.

Im Schritt S16 werden die Verbindungsstellen der einzelnen Nachbildungsmakrosegmente aneinander angeglichen, wie es in Fig. 3 dargestellt ist. Hierbei werden die Frequenzen links $f_l$ und rechts $f_r$ von den Verbindungsstellen V einander angepaßt, wobei die Endbereiche der Nachbildungsmakrosegmente vorzugsweise derart verändert werden, daß die Frequenzen $f_l$ und $f_r$ den gleichen Wert besitzen. Vorzugsweise kann im Bereich der Verbindungsstelle der Übergang auch geglättet und/oder stetig gemacht werden.

Nachdem für alle linguistisch phonetischen Einheiten des Textes die Nachbildungsmakrosegmente der Grundfrequenz erstellt und zusammengesetzt worden sind, wird das Unterprogramm beendet und der Programmablauf geht zurück zum Hauptprogramm (S17).

Mit dem erfindungsgemäßen Verfahren kann somit ein Verlauf einer Grundfrequenz erzeugt werden, der der Grundfrequenz einer natürlichen Sprache sehr ähnlich ist, da mittels des neuronalen Netzwerkes einfach größere Kontextbereiche erfaßt und ausgewertet werden können (Makrostruktur) und zugleich mittels der in der Datenbasis gespeicherten Grundfrequenzsequenzen feinste Strukturen des Grundfrequenzverlaufes entsprechend der natürlichen Sprache erzeugt werden können (Mikrostruktur). Hierdurch wird eine Sprachausgabe mit einem wesentlich natürlicheren Klang als bei bisher bekannten Verfahren ermöglicht.

Die Erfindung ist oben anhand eines Ausführungsbeispiels näher erläutert worden. Die Erfindung ist jedoch nicht auf das

konkrete Ausführungsbeispiel beschränkt, sondern im Rahmen der Erfindung sind unterschiedlichste Abwandlungen möglich. So kann z.B. die Reihenfolge, wann die Grundfrequenzsequenzen aus der Datenbasis und wann das neuronale Netzwerk das Vorgabemakrosegment erstellt, variiert werden. Es ist z.B. auch möglich, daß zunächst für alle phonetisch linguistischen Einheiten Vorgabemakrosegmente erzeugt werden und dann erst die einzelnen Grundfrequenzsequenzen ausgelesen, kombiniert, bewertet und ausgewählt werden. Im Rahmen der Erfindung können auch unterschiedlichste Kostenfunktionen angewandt werden, solange sie eine Abweichung zwischen einem Vorgabemakrosegment der Grundfrequenz und Mikrosegmente der Grundfrequenzen berücksichtigen. Das oben beschriebene Integral der lokalen Kostenfunktion kann aus numerischen Gründen auch als Summe dargestellt werden.

Patentansprüche

1. Verfahren zum Bestimmen des zeitlichen Verlaufs einer Grundfrequenz einer zu synthetisierenden Sprachausgabe, umfassend die Schritte:

Bestimmen von Vorgabemakrosegmenten der Grundfrequenz mittels eines neuronalen Netzwerkes, und

Bestimmen von Mikrosegmenten mittels in einer Datenbasis gespeicherten Grundfrequenzsequenzen, wobei die Grundfrequenzsequenzen derart aus der Datenbasis ausgewählt werden, daß durch die aufeinanderfolgenden Grundfrequenzsequenzen das jeweilige Vorgabemakrosegment mit möglichst geringer Abweichung nachgebildet wird.

2. Verfahren nach Anspruch 1,
d a d u r c h g e k e n n z e i c h n e t,
daß die Vorgabemakrosegmente einen Zeitbereich abdecken, der einer phonetisch linguistischen Einheit der Sprache, wie z.B. einer Phrase, einem Wort oder einer Silbe, entspricht.

3. Verfahren nach Anspruch 1 oder 2,
d a d u r c h g e k e n n z e i c h n e t,
daß die Grundfrequenzsequenzen der Mikrosegmente die Grundfrequenzen jeweils eines Phonems darstellen.

4. Verfahren nach einem der Ansprüche 1 bis 3,
d a d u r c h g e k e n n z e i c h n e t,
daß die Grundfrequenzsequenzen der Mikrosegmente, die innerhalb eines zeitlichen Bereiches eines der Vorgabemakrosegmente liegen, zu einem Nachbildungsmakrosegment zusammengesetzt werden, wobei die Abweichung des Nachbildungsmakrosegments zum jeweiligen Vorgabemakrosegment ermittelt wird, und die Grundfrequenzsequenzen derart optimiert werden, daß die Abweichung möglichst gering ist.

5. Verfahren nach Anspruch 4,
d a d u r c h g e k e n n z e i c h n e t,

daß für die einzelnen Mikrosegmente jeweils mehrere Grundfrequenzsequenzen ausgewählt werden können, wobei diejenigen Kombinationen von Grundfrequenzsequenzen ausgewählt werden, die die geringste Abweichung zwischen dem jeweiligen Nachbildungsmakrosegment und dem jeweiligen Vorgabemakrosegment ergeben.

6. Verfahren nach Anspruch 4 oder 5,
d a d u r c h g e k e n n z e i c h n e t,
daß die Abweichung zwischen dem Nachbildungsmakrosegment und dem Vorgabemakrosegment mittels einer Kostenfunktion ermittelt wird, die derart gewichtet ist, daß bei geringen Abweichungen von der Grundfrequenz des Vorgabemakrosegments lediglich eine kleine Abweichung ermittelt wird, wobei bei Überschreiten vorbestimmter Grenzfrequenzdifferenzen die ermittelten Abweichungen stark bis zum Erreichen eines Sättigungswertes ansteigen.

7. Verfahren nach einem der Ansprüche 4 bis 6,
d a d u r c h g e k e n n z e i c h n e t,
daß die Abweichung zwischen dem Nachbildungsmakrosegment und dem Vorgabemakrosegment mittels einer Kostenfunktion ermittelt wird, mit der eine Vielzahl von über die Makrosegmente verteilt angeordnete Abweichungen bewertet werden, wobei die Abweichung desto schwächer gewichtet werden, je näher sie am Rand einer Silbe angeordnet sind.

8. Verfahren nach einem der Ansprüche 4 bis 7,
d a d u r c h g e k e n n z e i c h n e t,
daß beim Auswählen der Grundfrequenzsequenzen die einzelnen Grundfrequenzsequenzen mit den hierzu jeweils nachfolgenden bzw. vorhergehenden Grundfrequenzsequenzen nach vorbestimmten Kriterien abgestimmt werden, und lediglich Kombinationen von Grundfrequenzsequenzen zum Zusammensetzen zu einem Nachbildungsmakrosegment zugelassen werden, die die Kriterien erfüllen.

9. Verfahren nach Anspruch 8,
dadurch gekennzeichnet,
daß die Beurteilung benachbarter Grundfrequenzsequenzen mittels einer Kostenfunktion erfolgt, die einen zu minimierenden Ausgabewert für eine Verbindungsstelle der Grundfrequenzsequenzen benachbarter Grundfrequenzsequenzen erzeugt, der desto größer ist, je größer die Differenz am Ende der vorhergehenden Grundfrequenzsequenz zur Frequenz am Anfang der nachfolgenden Grundfrequenzsequenz ist.

10. Verfahren nach Anspruch 9,
dadurch gekennzeichnet,
daß die der Ausgabewert desto schwächer gewichtet wird, je näher die jeweilige Verbindungsstelle am Rand einer Silbe angeordnet ist.

11. Verfahren nach einem der Ansprüche 1 bis 10,
dadurch gekennzeichnet,
daß die einzelnen Makrosegmente mitteinander verkettet werden, wobei an den Verbindungsstellen der Makrosegmente die Grundfrequenzen aneinander angepaßt werden.

12. Verfahren nach einem der Ansprüche 1 bis 11,
dadurch gekennzeichnet,
daß die neuronalen Netzwerke die Vorgabesegmente für einen vorbestimmten Abschnitt eines Textes auf Grundlage dieses Textabschnittes und eines diesem Textabschnitt vorausgehenden und/oder nachfolgenden Textabschnittes bestimmen.

13. Verfahren zum Synthetisieren von Sprache, bei dem ein Text in eine Folge akustischer Signale gewandelt wird, umfassend folgende Schritte:
Wandeln des Textes in eine Folge von Phonemen,
Erzeugen einer Betonungsstruktur,
Bestimmen der Dauer der einzelnen Phoneme,
Bestimmen des zeitlichen Verlaufs einer Grundfrequenz nach dem Verfahren gemäß einem der Ansprüche 1 bis 12,

Erzeugen der die Sprache darstellenden akustischen Signale auf Grundlage der ermittelten Folge von Phonemen und der ermittelten Grundfrequenz.

FIG 1A

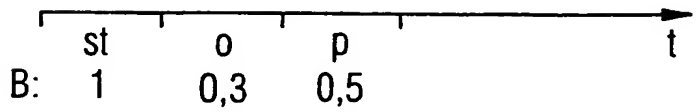



FIG 1B

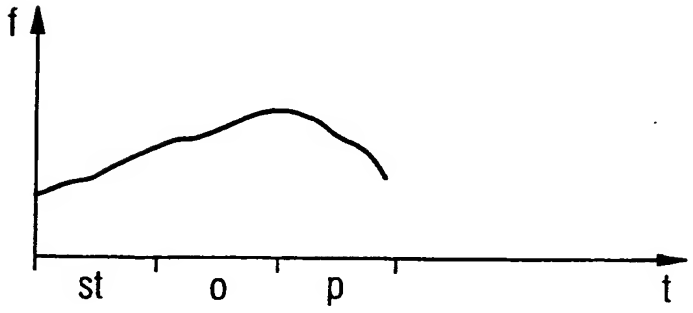



FIG 1C

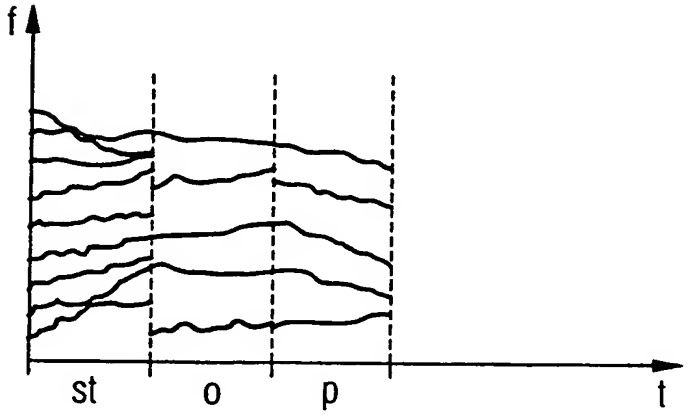



FIG 1D

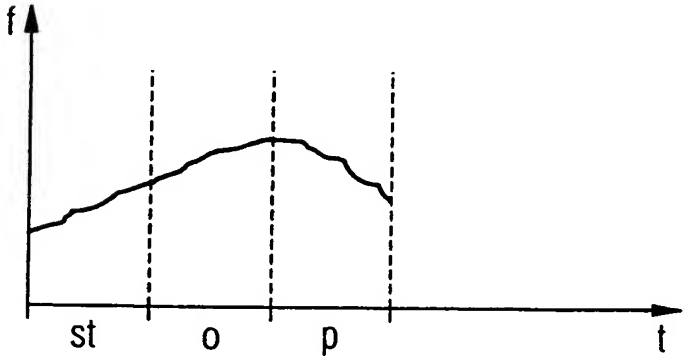


FIG 2

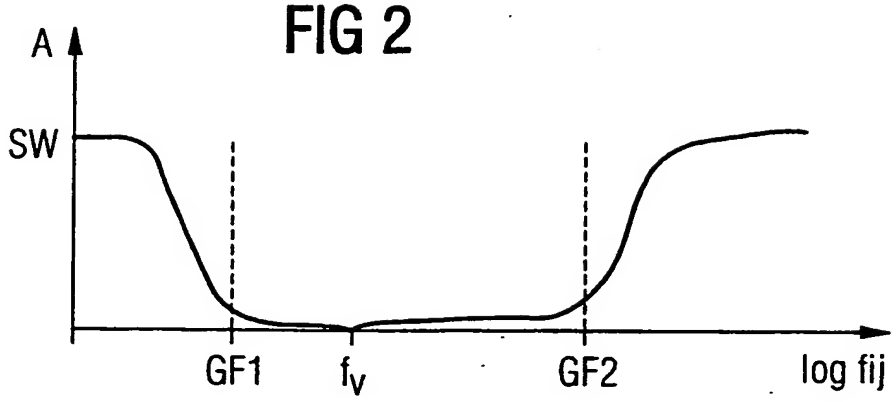



FIG 3

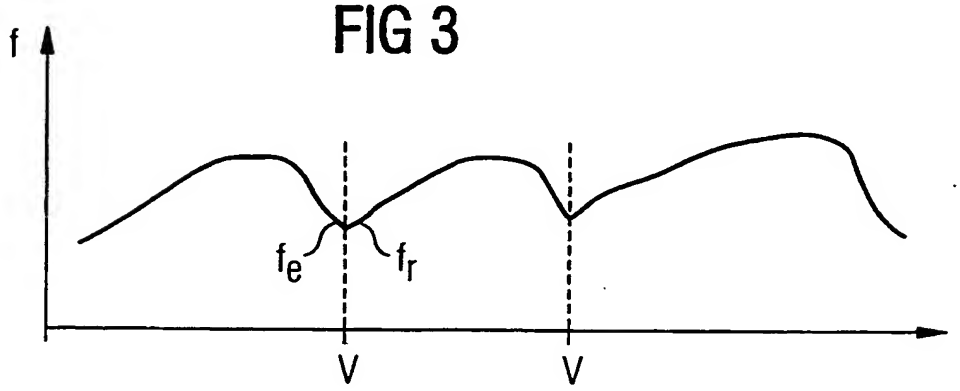



FIG 4

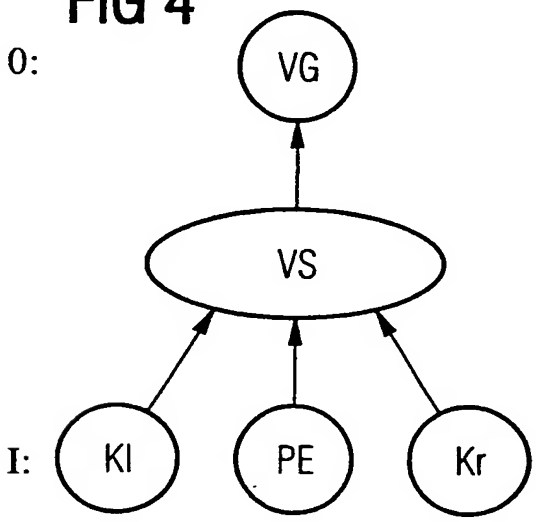


FIG 5

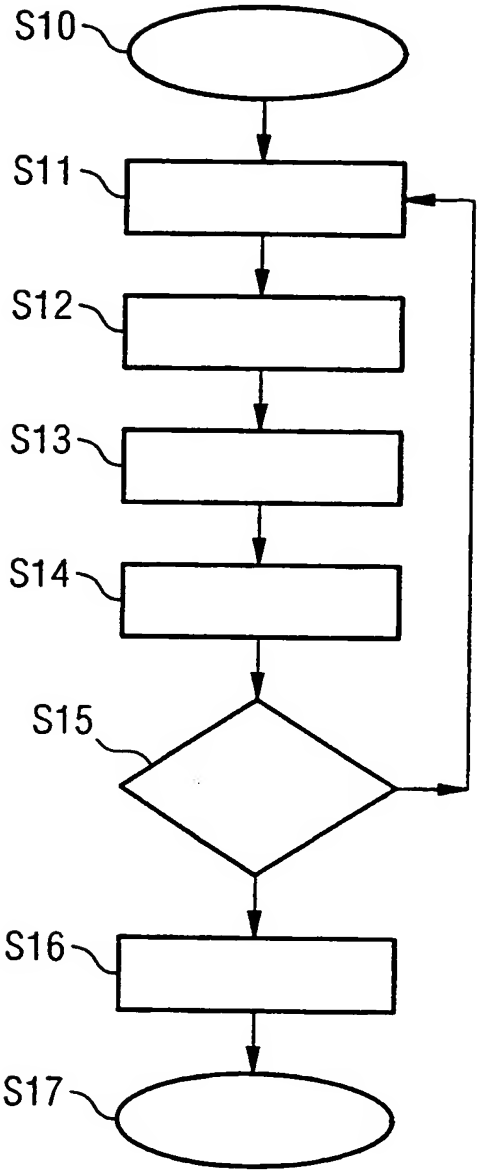



FIG 6